A fully automatic deep learning algorithm to segment Rectal Cancer on MR images: a multi-center study

A fully automatic deep learning algorithm to segment rectal Cancer on MR images: a multi-center study*

Jovana Panic, Arianna Defeudis, Simone Mazzetti, Samanta Rosati, Giuliana Giannetto, Monica Micilotta, Lorenzo Vassallo, Marco Gatti, Daniele Regge, Gabriella Balestra, *Member, IEEE*, Valentina Giannini, *Member, IEEE*

***Abstract*— The aim of the study is to present and tune a fully automatic deep learning algorithm to segment colorectal cancers (CRC) on MR images, based on a U-Net structure. It is a multicenter study, including 3 different Italian institutions, that used 4 different MRI scanners. Two of them were used for training and tuning the systems, while the other two for the validation. The implemented algorithm consists of a pre-processing step to normalize and to highlight the tumoral area, followed by the CRC segmentation using different U-net structures. Automatic masks were compared with manual segmentations performed by three experienced radiologists, one at each center. The two best performing systems (called *mdl2* and *mdl3*), obtained a median Dice Similarity Coefficient of 0.68(*mdl2*) - 0.69(*mdl3*), precision of 0.75(*mdl2*) - 0.71(*mdl3*), and recall of 0.69(*mdl2*) - 0.73(*mdl3*) on the validation set. Both systems reached high detection rates, 0.98 and 0.95, respectively, on the validation set. These encouraging results, if confirmed on larger dataset, might improve the management of patients with CRC, since it can be used as a fast and precise tool for further radiomics analyses.**

***Clinical Relevance*— To provide a reliable tool able to automatically segment CRC tumors that can be used as first step in future radiomics studies aimed at predicting response to chemotherapy and personalizing treatment.**

I. Introduction

Colorectal cancer (CRC) is the third tumor in terms of incidence and mortality, and 60% of CRC are diagnosed as Locally Advanced Rectal Cancer (LARC)[1]. The recommended treatment is neoadjuvant chemoradiotherapy (nCRT) followed by Total Mesorectal Excision (TME)[2]. Despite the advantages shown by nCRT, patients' response varies widely, ranging from completely response (up to 20% of cases) to no response or tumor progression [3]. Artificial intelligence (AI) has shown promises in the development of radiomics signature, based on Magnetic Resonance Imaging (MRI), that can predict patient's response to nCRT, thus allowing more personalized treatments [4]–[7]. Despite the promising results, the translation of this approaches into clinical practice is still limited by many reasons, including the lack of automatic segmentation methods. Indeed, both manual and semi-automatic segmentations methods have two main drawbacks: they are a time-consuming task, that has to be regarded as prohibitive when very large databases are evaluated, and they lead to a high inter-reader variability that can strongly impact on the performance of predictive tools [6]. Therefore, developing automatic segmentation methods is of key importance to realize robust tools that can be effectively used in the clinical practice. In the last few years, Deep Learning (DL) algorithms have been used in the medical imaging field to segment and detect anatomical structures [8], [9]. More recently, the U-Net architecture [10] has been presented to overcome some limitations of previously developed structures, i.e., Fully Convolutional Networks (FCNs) and Convolutional Neural Networks (CNNs). The main advantage of the U-Net structure is the absence of the fully connected layer, replaced by the up-sampling layer and the deconvolutional layer, which allow to obtain a probability score map with the same size of the input, classifying each pixel instead the whole image [10]. To the best of our knowledge, only few studies used the U-Net to automatically localize and segment LARC on MR images [11]–[13]. However, all these methods require an initial manual crop of the image to delimit the region of interest. Moreover, none of them was validated on an external dataset.

In this study, we developed and tuned a fully automatic U-Net architecture to segment LARC on MRI that was validated on an external dataset composed of images acquired on a center not involved in the training phase.

II. Materials and Methods

A. Patients and reference standard

Patients with proven locally LARC were retrospectively collected from three different institutions, with four different scanners (GE Health Care GDx Signa Excite and GE Health

* This research was funded by AIRC 5xmille Special Program Molecular Clinical Oncology - Ref. 9970, FPRC 5xmille 2013 Ministero della Salute, and FPRC 5xmille 2015 Ministero della Salute (STRATEGY), Fondazione AIRC under 5 per Mille 2018—ID. 21091 program—P.I.Bardelli Alberto, G.L. Regge Daniele.

J. P., S.R, G.B are with the Polytechnic of Turin, Department of Electronics and Telecommunications, Turin, Italy (corresponding author; e-mail: jovana.panic@ polito.it).

A.D., S.M, D.R, V.G are with University of Turin, Department of Surgical Science, Turin, Italy and Candiolo Cancer Institute FPO-IRCCS, Italy.

G. G. is with Candiolo Cancer Institute FPO-IRCCS, Italy.

M. M. is with A. O. Ordine Mauriziano, Turin, Italy.

L. V. is with Radiology Unit, SS Annunziata Savigliano Hospital, Cuneo, Italy.

M. G. is with University of Turin, Department of Surgical Sciences - Radiology Unit, Turin, Italy.

Care Optima MR450w A, Philips Ingena for center B and Philips Achieva for center C). All patients underwent multiparametric (mp)MRI before nCRT after October 2010, including T2 weighted (T2w) and Diffusion-Weighted Imaging (DWI) sequences according to MRI guidelines for reporting rectal cancer staging [14]. All tumors were manually segmented on T2w images by a resident radiologist per each center and revised by a second radiologist with more than 10-year experience in mpMR imaging. This was a multi-center retrospective study approved by the institutional review boards (IRBs) in each institution, with a waiver for requirement of informed consent as de-identified data were used.

B. Pre-processing

The pre-processing phase consists of three steps: the evaluation of the Apparent Diffusion Coefficient (ADC), the cropping and the normalization of the images.

First, the ADC is calculated from the DWI sequences of each patient according to the mono-exponential equation. The second step consists in the automatic crop of the images around the bounding box containing the tumor, to reduce the amount of irrelevant information and to minimize the differences among patients during the normalization step, as previously described [15]. Finally, all cropped images were normalized to account for differences arose from different scanners. Both T2w and ADC sequences are standardized using the z-score normalization.

C. Construction of the Training set

The proposed system has been constructed using patients from scanners A.1 and C, pooled together to obtain the construction set, that was further divided into a training and a test set. The latter was used for the tuning process. Finally, the system was externally validated on the sequences from scanners A.2 and B. To obtain a balanced training set, we collected the same number of slices with and without tumor. All tumoral slices were included, while the non-tumoral ones were chosen randomly among all slices of all patients. Due to the high dependency of the performances of the DL algorithms to the training dataset, we developed and compared two different procedures to build the training set:

- Random sampling (*tr_rnd*): this method is based on a random selection of 70% of the patients from A.1 and 70% from C, while all the remaining cases are included in the testing set.
- Sampling based on clustering (*tr_dnd*): this approach is based on an agglomerative hierarchical clustering method that organizes data in a hierarchical tree (called dendrogram) based on a proximity measure. Then, the final clusters are obtained by cutting the tree at a certain level [16]. To apply this approach, first we extracted the following 20 features for each patient of A.1 and C: mean, standard deviation, median, 25th and 75th percentile of both the LARC volume and the whole cropped volume in the T2w and ADC sequences. Then, we applied the hierarchical clustering to these patients, and we cut the tree to create two clusters. The training set was created by randomly collecting the same number of patients (70% of the less numerous cluster) from both clusters. The discarded patients of the clusters were included in the testing set.

D. U-Net

Two different hyper-parameters have been analyzed to observe how they affected the performances of the system: the loss function and the number of descending levels.

- Loss Function: we analysed the Binary Crossentropy (BC), which suits binary classification tasks [17], and a custom loss function (CL), which overcomes the issues related to the class imbalance, since there is a higher number of non-tumoral voxels against tumoral ones. It was implemented by merging the BC and the Dice Loss:

$$\mathrm{CL} = -\frac{1}{\mathrm{N}}\sum \mathrm{y_i}\cdot\log\big(\mathrm{p(y_i)}\big) + 1 - \frac{2\sum \mathrm{y_i}\cdot\mathrm{p(y_i)}}{\sum \mathrm{y_i}+\sum \mathrm{p(y_i)}}, \quad (1)$$

where y_i is the *i*-th label and $p(y_i)$ is the predicted probability of the sample to belong to the *i*-th label class.

- Number of descending levels: it defines the complexity of the features evaluated by the U-Net, strongly affecting the learning process. In our study we considered U-Net structures with 3, 4, and 5 descending levels.

All networks were implemented in Python (v. 3.7.4), using the Tensorflow (v. 2.2.0) library, with the Adam optimizer [18] and a starting learning rate value of 0.001, β1 of 0.9 and β2 of 0.999.

The tuning process of the different training sets, loss functions and number of descending levels lead us to develop and evaluate 12 U-Net systems. The best configurations on the test set were then validated on the validation set.

E. Post-processing

Finally, the mask was binarized using the Otsu's threshold of the predicted mask, and to reduce the false positive elements detected by the U-Net, characterized by spatially connected areas on less than 3 slices.

F. Validation

Once we developed all networks, we select the networks which presented the higher DSC, Pr and Re (see following section) respectively on the testing set. We also set the condition that all three parameters must have been higher than 0.6, to exclude the models which over- or under-segmented the tumoral volume. Those models were then validated on the external validation dataset, and their performances were analyzed using the same parameters as for the testing set.

G. Statistical analysis

For this study the network performances were evaluated using the following parameters:

- Dice Similarity Coefficient (DSC):

$$\mathrm{DSC} = \frac{2\mathrm{TP}}{\mathrm{FP}+2\mathrm{TP}+\mathrm{FN}}, \quad (2)$$

where TP is True Positive voxel, FP is False Positive voxel, and FN is False Negative voxel.

- Precision (Pr):

$$\mathrm{Pr} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \quad (3)$$

where TP is True Positive voxel, FP is False Positive voxel.

- Recall (Re):

$$\mathrm{Re} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}, \quad (4)$$

where TP is True Positive voxel, FN is False Negative voxel. A tumor was defined FN if its DSC was lower than 0.2, while the Detection Rate (DR) was defined as the percentage of correctly detected tumors (DSC>0.2), per each model.

III. Results

A. Patients

100 patients (61 men and 39 women) were retrospectively collected, having an average age of 64 years (range 34-86). Fifty-eight patients were included in the construction set and the remaining 42 were used as validation set. *Tr_rnd* was composed of 41/58 patients, resulting in 222 tumoral and 222 non tumoral 256x256 slices. *Tr_dend* showed a slightly diffferent number of patients due to the agglomerative hierarchical clustering method, i.e., 36/58 patients (211 tumoral and 211 non tumoral 256x256 slices).

B. Tuning of parameters using the construction set

Table I shows the performances of the models with different descending levels, considering all combinations of training sets and loss functions. In particular, the Pr results related to the lowest number of descending levels (n. 3) do not satisfy the condition of Pr>0.60 (Pr=0.58, 0.56, 0.59), except for the *3lv_BC std_rnd* model (Pr=061*)*. On the other hand, the Pr values related to higher number of descending levels (n. 4 and n. 5), are comparable to each other. There are two models which do not meet the condition of Re>0.60: *4lv_CL std_dend* (Re=0.55) and *5lv_CL std_rnd* (Re=0.58). It is possible to notice that the values of Pr and Re are strongly related to the combination of training set and loss function, while the DSC values are comparable considering the different combinations.

In conclusion, we selected the three different models that reaches the highest DSC, Pr and Re:

- *5lv_CL~std_dend* – from now on *mdl1*: DSC=0.73, Pr=0.69, Re=0.78;
- *4lv_CL~std_rnd* – from now on *mdl2*: DSC=0.64, Pr=0.74, Re=0.62;
- *4lv_BC~std_dend* – from now on *mdl3*: DSC=0.66, Pr=0.67, Re=0.78.

C. Validation set

Table II shows the performances of the three best models on the validation set. *mdl1* obtained the highest Re, however, performance of *mdl2* and *mdl3* were higher, if we consider the DR. Indeed, sensitivity in the validation set for *mdl1* is 88% with 5/42 FN, for *mdl2* is 98% with 1/42 FN (2.4%), and for *mdl3* is 95% with 2/42 FN (4.8%). Fig. 1 shows the differences between the segmentations obtained by the three models for 2 patients of the validation set: in patient 73 (Fig 1.a) most of the tumor is correctly classified, but there are some FP areas, in particular by *mdl1*. Fig. 1.b (patient 208) show several misclassified areas: the first one is not detected by *mdl1*, while *mdl2* reaches DSC=0.31 and *mdl3* DSC=0.12 on the volume. The main reason is probably due to the different pixel intensities on both T2w and ADC sequences. Moreover, it is wrongly cropped after the pre-processing step. In the validation set there are 8/42 cases where the tumoral volume is not correctly cropped, but at least 95% of the tumoral voxels are included.

TABLE I. PERFORMANCES ON THE TESTING SET

Coefficients / U-Net	***DSC*** Median (IQR)	***Pr*** Median (IQR)	***Re*** Median (IQR)
3lv_BC std_rnd	0.64 (0.52 – 0.71)	0.61 (0.49 – 0.74)	0.67 (0.52 – 0.87)
3lv_CL std_rnd	0.63 (0.44 – 0.71)	0.58 (0.49 – 0.76)	0.64 (0.41 – 0.83)
3lv_BC std_dend	0.62 (0.54 – 0.70)	0.56 (0.46 – 0.69)	0.78 (0.60 – 0.91)
3lv_CL std_dend	0.65 (0.56 – 0.70)	0.59 (0.51 – 0.74)	0.68 (0.51 – 0.80)
4lv_BC std_rnd	0.60 (0.49 – 0.72)	0.63 (0.50 – 0.77)	0.64 (0.48 – 0.83)
4lv_CL std_rnd	0.64 (0.47 – 0.72)	**0.74** (0.66 – 0.82)	0.62 (0.37 – 0.72)
4lv_BC std_dend	0.66 (0.63 – 0.76)	0.67 (0.55 – 0.75)	**0.78** (0.63 – 0.83)
4lv_CL std_dend	0.63 (0.48 – 0.72)	0.75 (0.63 – 0.82)	0.55 (0.39 – 0.76)
5lv_BC std_rnd	0.66 (0.60 – 0.72)	0.65 (0.60 – 0.77)	0.64 (0.61 – 0.75)
5lv_CL std_rnd	0.64 (0.48 – 0.69)	0.70 (0.59 – 0.78)	0.58 (0.39 – 0.75)
5lv_BC std_dend	0.68 (0.61 – 0.74)	0.70 (0.60 – 0.80)	0.70 (0.62 – 0.84)
5lv_CL std_dend	**0.73** (0.66 – 0.75)	0.69 (0.59 – 0.78)	0.78 (0.70 – 0.90)

BC = Binary Crossentropy, CL = Custom Loss, dend = training and testing obtained by the hierarchical clustering method, lv. = descending levels, std = standardized sequences, values underlined = under the threshold (0.60), values in bold = highest values obtained by the models for each parameter, green rows = best performing models on the testing set.

IV. Discussion

In this study, we developed and tuned several U-Net based systems for the automatic segmentation of LARC on mpMRI. Two of them reached promising results on both the construction and the validation set: *mdl2* with median DSC of 0.68, Pr of 0.75 and Re of 0.69, and *mdl3* with DSC of 0.69, Pr of 0.71 and Re of 0.73. Soomro et al. [11] presented a densely interconnected 3D-FCN based model, which reached good performances on the test sets, showing a DSC from 0.83 to 0.93. Li et al.[12] and used a U-Net algorithm, as we did, obtaining higher DSC, i.e., 0.74 and 0.98, respectively, vs 0.68(*mld2*) and 0.69(*mdl3).* However, most of these methods

TABLE II. PERFORMANCES ON THE VALIDATION SET

Coefficients / U-Net	***DSC*** Median (IQR)	***Pr*** Median (IQR)	***Re*** Median (IQR)	***Detection Rate***
5lv_CL_std_dend	0.62 (0.43 – 0.73)	0.60 (0.50 – 0.74)	0.75 (0.42 – 0.89)	0.88 (37/42)
4lv_CL_std_rnd	0.68 (0.52 – 0.77)	0.75 (0.62 – 0.82)	0.69 (0.49 – 0.75)	0.98 (41/42)
4lv_BC_std_dend	0.69 (0.52 – 0.75)	0.71 (0.55 – 0.79)	0.73 (0.55 – 0.84)	0.95 (40/42)

BC = Binary Crossentropy, CL = Custom Loss, dend = training and testing obtained by the dendrogram method, lv. = descending levels.

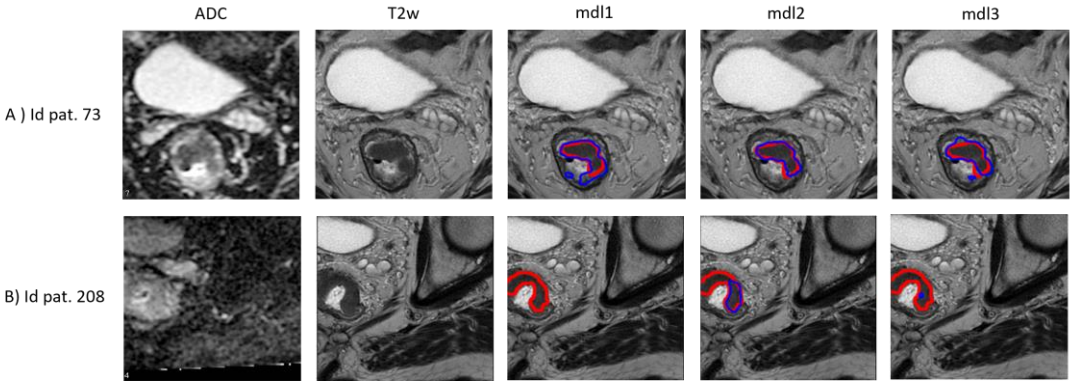

Figure 2. Validation examples of correctly detected tumor (A) and a FN example (B) wrongly detected by *mdl1*, *mdl2* and *mdl3*. The first and second columns show the ADC and T2w sequences each, while from the third to the fourth columns the predictions of *mdl1*, *mdl2* and *mdl3* respectively. The manual segmentation is contoured by the red line, while the network's predictions with the blue line.

were not fully automatic and any of them was validated on an external dataset. The latter is a key point, when developing robust radiomics CAD systems for clinical use. Indeed, it is strongly recommended to use multi-centric dataset, acquired with different acquisition scanners and protocols [19].

To the best of our knowledge, only Knuth et al. [13] presented a U-Net trained on mpMR images from two different centers. In our study we obtained higher DSC (0.68(*mld2*) and 0.69(*mdl3)* vs 0.59). The second strength of our method relies on the fact that we focused our analyses also on the optimization of different parameters of the U-Net architecture, showing the need to fine-tune DL nets.

This study has also some limitations. First, the high dependency of the cropping phase in pre-processing to the DWI image quality and acquisition parameters, and the use of the ADC sequences to perform this step. The latter can be an issue for those centers that do not acquire DWI sequences during their clinical protocol. Second, the size of the construction set (n=58) should be increased by including more centers and possibly developing a prospective clinical trial. Third, we did not train the U-Net by combining information from both 2D and 3D volumes.

In conclusion, the developed U-Net based systems present promising performances on an external validation set, being able to identify the tumoral volumes on MR images acquired with different scanners and acquisition protocols. Hopefully, these systems may lead to future development and inclusion of automatic detection and prediction systems of CRC in clinical pathways.

REFERENCES

[1] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer Statistics, 2021," *CA. Cancer J. Clin.*, vol. 71, no. 1, pp. 7–33, Jan. 2021.

[2] A. B. Benson *et al.*, "Rectal cancer, version 6.2020: Featured updates to the NCCN guidelines," *JNCCN Journal of the National Comprehensive Cancer Network*, vol. 18, no. 7. pp. 807–815, Jul-2020.

[3] I. J. Park *et al.*, "Neoadjuvant treatment response as an early response indicator for patients with rectal cancer," *J. Clin. Oncol.*, vol. 30, no. 15, pp. 1770–1776, 2012.

[4] P. Bulens *et al.*, "Predicting the tumor response to chemoradiotherapy for rectal cancer: Model development and external validation using MRI radiomics," *Radiother. Oncol.*, vol. 142, pp. 246–252, 2020.

[5] V. Giannini *et al.*, "Predicting locally advanced rectal cancer response to neoadjuvant therapy with 18 F-FDG PET and MRI radiomics features," *Eur. J. Nucl. Med. Mol. Imaging*, vol. 46, no. 4, pp. 878–888, Apr. 2019.

[6] F. Coppola *et al.*, "Radiomics and magnetic resonance imaging of rectal cancer: From engineering to clinical practice," *Diagnostics*, vol. 11, no. 5. p. 756, 2021.

[7] V. Giannini, S. Rosati, C. Castagneri, L. Martincich, D. Regge, and G. Balestra, "Radiomics for pretreatment prediction of pathological response to neoadjuvant therapy using magnetic resonance imaging: Influence of feature selection," in *Proceedings - International Symposium on Biomedical Imaging*, 2018, vol. 2018-April, pp. 285–288.

[8] X. Fang, S. Xu, B. J. Wood, and P. Yan, "Deep learning-based liver segmentation for fusion-guided intervention," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 15, no. 6, pp. 963–972, 2020.

[9] B. Lee, N. Yamanakkanavar, and J. Y. Choi, "Automatic segmentation of brain MRI using a novel patch-wise U-net deep architecture," *PLoS One*, vol. 15, no. 8 August, pp. 1–20, 2020.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015, vol. 9351, pp. 234–241.

[11] M. H. Soomro *et al.*, "Automated segmentation of colorectal tumor in 3D MRI Using 3D multiscale densely connected convolutional neural network," *J. Healthc. Eng.*, vol. 2019, 2019.

[12] D. Li, X. Chu, Y. Cui, J. Zhao, K. Zhang, and X. Yang, "Improved U-Net based on contour prediction for efficient segmentation of rectal cancer," *Comput. Methods Programs Biomed.*, vol. 213, p. 106493, 2022.

[13] F. Knuth *et al.*, "MRI-based automatic segmentation of rectal cancer using 2D U-Net on two independent cohorts," *Acta Oncol. (Madr).*, vol. 0, no. 0, pp. 1–9, 2021.

[14] V. Granata *et al.*, "Structured reporting of lung cancer staging: A consensus proposal," *Diagnostics*, vol. 11, no. 9, pp. 1–33, 2021.

[15] J. Panic *et al.*, "A Convolutional Neural Network based system for Colorectal cancer segmentation on MRI images," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2020, vol. 2020-July, pp. 1675–1678.

[16] R. O. Duda, P. E. Hart, D. G. Stork, and J. Wiley, "Pattern Classification All materials in these slides were taken from Pattern Classification (2nd ed)," no. April, 2016.

[17]. Usha Ruby Dr.A, "Binary cross entropy with deep learning technique for Image classification," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 4, pp. 5393–5397, 2020.

[18] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[19] U. Bağci, J. K. Udupa, and L. Bai, "The role of intensity standardization in medical image registration," *Pattern Recognition Letters*, vol. 31, no. 4. pp. 315–323, 2010.